002.513.5.001:025.4

Teoretyczne problemy systemów wyszukiwania informacji

ZDZISŁAW PAWLAK
Centrum obliczeniowe PAN, Warszawa

Komputerowa realizacja systemów wyszukiwania informacji wyłoniła szereg problemów zarówno teoretycznej jak i praktycznej natury. Powstało wiele metod pozwalających na szybkie i sprawne organizowanie procesu wyszukiwania, a także wiele teorii matematycznych próbujących sprecyzować zasady organizowania takich procesów oraz ich podstawowe własności. Jak dotąd brak jednakże prób zbudowania konsekwentnej, jednolitej teorii, precyzującej podstawowe pojęcia systemów wyszukiwania informacji. Wydaje się, że podstawy takie są niezbędne nie tylko dla lepszego rozumienia struktury systemów wyszukiwania informacji — a więc i dla dydaktyki — ale również dla prowadzenia dalszych badań w tej dziedzinie.

W niniejszym artykule chciałbym przedstawić próbę sprecyzowania pewnych pojęć związanych z wyszukiwaniem informacji, zdając sobie doskonale sprawę, że jest ona daleka od doskonałości. Cechą charakterystyczną tej próby jest przede wszystkim konsekwentne wprowadzenie języka informacyjnego i traktowanie go podobnie jak to się czyni z językami w logice, a ostatnio również z językami programowania, tj. określa się ich składnię i semantykę. W wyniku takiego ujęcia problemu powstała nowa metoda organizowania systemów wyszukiwania informacji — istotnie różna od metod stosowanych do tej pory. Pozwala ona na znacznie szybsze organizowanie procesu wyszukiwania niż metody dotychczasowe, a co więcej szybkość tę można w pewnym sensie regulować, tzn. istnieje możliwość takiego zaprojektowania systemu aby odpowiedź na określone pytania nie przekraczała z góry ustalonego czasu. Istnieją naturalnie pewne ograniczenia i czas ten nie może być dowolnie krótki, minimalny czas odpowiedzi zależy bowiem zarówno od samej maszyny jak i od organizacji systemu. Jest on jednakże na ogół wielokrotnie krótszy niż przy zastosowaniu metod klasycznych.

Innym faktem uzyskanym dzięki omawianym rozważaniom jest dokładne zbadanie roli języka w systemach wyszukiwania informacji. Chodzi tu o to, że w ogólnym przypadku język nie pozwala na wyrażanie dowolnych własności dotyczących opisywanych obiektów, istnieją tu bowiem pewne ograniczenia. Fakt ten ma być może większe znaczenie poznawcze aniżeli praktyczne, jednakże należy sobie zdawać sprawę również w przypadkach konkretnych, z granic możliwości systemu.

Bardziej szczegółowe informacje nt. poruszonych tu problemów znajdzie czytelnik w pracy — W. Marek Z. Pawlak: Information Retrieval Systems — Mathematical Foundations, Prace COPAN, 1974, nr 149.

1. System informacyjny

W każdym systemie wyszukiwania informacji mamy do czynienia z pewnymi obiektami, którymi mogą być czasopisma, książki, wycinki prasowe etc. Obiekty te są charakteryzowane poprzez określenie interesujących nas własności tych obiektów. W przypadku wymienionych wyżej dokumentów własnościami takimi mogą być np.: nazwisko autora dokumentu, język i data publikacji, przedmiot etc. Powstaje od razu pytanie czy taka charakterystyka obiektu (poprzez podanie jego wyróżnionych własności) jest dobra?, czy jest jedyna?, jakie są ewentualnie inne możliwe sposoby definiowania przedmiotów, które nas interesują?

Wiadomo, że nie jest to metoda najlepsza. Powodów tego jest wiele, nie będziemy jednak o nich szerzej mówić, gdyż nie jest to tematem tych rozważań. Zauważmy tylko rzecz powszechnie znaną, iż obiekt, którym się interesujemy może swe własności zmieniać w czasie (np. kolor włosów — informacja w dowodzie osobistym), tak że raz podane cechy mogą obowiązywać tylko przez określony okres czasu — bądź też nie możemy jakiejś własności ściślej sprecyzować. Mimo wielu wad tego sposobu charakteryzowania obiektów jest on obecnie podstawą systemów wyszukiwania informacji i w naszych rozważaniach przyjmiemy go jako punkt wyjścia.

Przyjmiemy więc, że w każdym systemie wyszukiwania informacji mamy do czynienia z pewnymi obiektami, które charakteryzujemy poprzez poddanie ich własności. Inaczej mówiąc z każdym obiektem

naszego systemu wiążemy wektor własności. Oczywiście każdy obiekt możemy w zasadzie charakteryzować poprzez różne własności, jednakże wymogi zarówno teorii jak i praktyki każą te wektory w jakiś sposób ujednolicić, tak aby wszystkie obiekty systemu były charakteryzowane jednakowo. Tak więc dla opisu obiektów w systemie wyszukiwania informacji wprowadzimy pojęcie systemu informacyjnego S, który określimy w następujący sposób:

$$S=\langle X, A, Q, \varrho\rangle$$

gdzie
X — zbiór obiektów S
A — zbiór atrybutów (własności) S
Q — zbiór deskryptorów S
ϱ — funkcja $\varrho: X\times A\to Q$

Funkcja ϱ charakteryzuje więc każdy obiekt naszego systemu przez przyporządkowanie mu określonych własności. Funkcję ϱ wygodnie jest przedstawić w postaci:

$$\varrho_x: A\to Q$$

traktując x jako parametr, lub też w postaci:

$$\overline{\varrho}: X\to F$$

gdzie F jest zbiorem wszystkich funkcji o argumentach w A i wartościach w Q.

Funkcje należące do zbioru F będziemy nazywać **informacjami**, zaś funkcje ϱ (a także ϱ_x i $\overline{\varrho}$) nazwiemy informacją o x w S.

Wprowadzone tu pojęcie „informacji" jest zgodne z potocznym rozumieniem tego słowa w sensie systemów wyszukiwania informacji (zwrócmy uwagę, że nie chodzi tu o informację w sensie Shannona). W myśl podanej definicji dowód osobisty jest informacją o osobie opisywanej w tym dowodzie, ankieta personalna jest informacją o osobie, której ona dotyczy, karta biblioteczna książki jest informacją o tejże książce.

Warto tu może dodać, że tak wprowadzone pojęcie systemu informacyjnego nie koniecznie musi być związane ze zbiorem dokumentów. Jako obiekty możemy przyjąć również stany pewnego przedmiotu i charakteryzować je poprzez wektor stanów. W tym wypadku otrzymamy również system informacyjny. Ma to o tyle sens, że rzeczywiste systemy informacyjne (komputerowe) odnoszą się nie tylko do zbiorów dokumentów, ale również w przypadku stosowania komputerów do sterowania procesami, do zbioru stanów sterowanych obiektów. Tak więc nasze rozważania mogą dotyczyć również tego rodzaju sytuacji.

Pojęcie informacji pozwala na wprowadzenie obiektów nierozróżnialnych w systemie tj. obiektów posiadających tę samą informację.

Powiemy że, $x\varepsilon$, X są równoważne (nierozróżnialne) w S, symbolicznie $x_S y$, lub przy ustalonym systemie S, $x\sim y$, gdy $\overline{\varrho}(x)=\overline{\varrho}(y)$.

Niech $X\varphi$ oznacza zbiór wszystkich obiektów należących do X posiadających informację $\varphi\varepsilon F$. Pojęcie informacji dzieli zbiór X na klasy spełniające następujące dwa warunki:

1. Jeżeli $\varphi\neq\psi$, to $X\varphi\cap X\psi=\Phi$
(Φ oznacza zbiór pusty)

$$\bigcup X\varphi=X$$

2. $\varphi\varepsilon F$

Inaczej mówiąc pojęcie informacji dzieli zbiór obiektów X na klasy rozłączne, których suma wyczerpuje cały zbiór X. Informacja wprowadza więc klasyfikację zbioru obiektów X w naszym systemie informacyjnym. Obiekty należące do tej samej klasy są w systemie nierozróżnialne.

2. Język systemu informacyjnego S

Z każdym systemem informacyjnym S wiążemy język formalny L_S, który będzie służył do opisywania obiektów (ściślej, klas obiektów) ze zbioru X oraz do opisywania pewnych ogólnych własności samego systemu S. Do opisywania klas obiektów posłuży zbiór termów (wyrażeń) języka — natomiast do opisywania własności systemu służą **formuły**, o których tu nie będziemy mówić. Termy można utożsamić więc z pytaniami. L_S jest więc językiem pytań. Natomiast zbiór odpowiadający pytaniu t będziemy nazywać odpowiedzią.

Termy języka L_S określone są następująco:

1. Stałe 0,1 są termami; każda para postaci $\langle a, q\rangle$ jest termem, gdzie a jest atrybutem zaś q deskryptorem.

2. Jeżeli t, t' są termami, to wyrażenia postaci $(t\cdot t)$, $(t+t')$, oraz $\sim(t)$ są również termami.

Termy są to więc wyrażenia powstałe z deskryptorów poprzez połączenie ich operacjami boolowskimi.

Dla określenia semantyki termów języka L_S wprowadzimy następującą funkcję:

$$\sigma: A\times Q\to_p(X)$$

gdzie $_p(X)$ oznacza zbiór wszystkich podzbiorów zbioru X.

Funkcją tą wygodniej posługiwać się traktując A jako parametr, tj. pisząc:

$$\sigma_a: Q\to_p(X)$$

Funkcja ta więc przyporządkowuje każdemu deskryptorowi zbiór obiektów posiadających własność wyrażoną przez dany deskryptor. Np. jeżeli jako deskryptor przyjmiemy nazwisko autora „Kowalski", w jakimś ustalonym systemie informacyjnym S, to wartością funkcji σ dla tego deskryptora będą wszystkie prace napisane przez Kowalskiego.

Semantyka (znaczenie) termu jest to funkcja, która każdemu termowi przyporządkowuje pewien podzbiór obiektów tj. $\sigma^*: T\to_p(X)$
T (jest zbiorem termów) jest określone w następujący sposób:

1.° $\sigma^*(a,q)=\sigma_a(q)$
2.° $\sigma^*(t\cdot t')=\sigma^*(t)\cap\sigma^*(t')$, $\sigma^*(t+t')=\sigma^*(t)\cup\sigma^*(t')$,
$\sigma^*(\sim t)=X-\sigma^*(t)$.

W ten sposób mając zadany term (pytanie) możemy obliczyć wszystkie obiekty, które ten term opisuje. Zbiór tych termów stanowi odpowiedź na pytanie.

3. Termy prymitywne i termy normalne (pytanie proste i normalne). Jeżeli z każdego atrybutu wybierzemy po jednym, dowolnym deskryptorze i połączymy je wszystkie symbolami iloczynu boolowskiego, to otrzymane w ten sposób wyrażenie nazwiemy termem (pytaniem) prostym albo **prymitywnym**. Każdy term prymitywny można uznać za językowy odpowiednik informacji. Każdej informacji odpowiada więc jeden term prymitywny i odwrotnie. Boolowską sumę dowolnych termów prymitywnych nazwiemy zaś termem

w postaci normalnej (lub pytaniem normalnym). Podstawowa własność języka L_S jest następująca: dla każdego termu t istnieje term t' w postaci normalnej, taki, że

$$\sigma^*(t)=\sigma^*(t')$$

Własność ta stanowi istotę naszych rozważań. Informuje ona mianowicie o tym, że odpowiedzi na oba pytania są identyczne.

4. Uwagi o implementacji

Semantyka języka L_S (a właściwie jego części dotyczącej termów) podawała od razu metodę liczenia odpowiedzi, (metoda ta jest znana pod nazwą list inwersyjnych). Po prostu dla znalezienia odpowiedzi na postawione pytanie należy odszukać w pamięci maszyny zbiory obiektów odpowiadających deskryptorom występującym w pytaniu, a następnie na tych zbiorach wykonać wskazane w termie działania teoriomnogościowe. O ile mamy w ten sposób „zapamiętane" obiekty w maszynie, że obiekty odpowiadające ustalonemu deskryptorowi są zmagazynowane razem, to obliczenie odpowiedzi jest stosunkowo proste.

Z podanej wyżej własności wynika jednakże jeszcze inna metoda liczenia odpowiedzi. Ponieważ każde pytanie ma odpowiadającą mu równoważną postać normalną, wynika więc z tego, że każda odpowiedź jest sumą pewnych odpowiedzi elementarnych. Odpowiedzi elementarne są to odpowiedzi na pytanie proste (termy prymitywne), które to termy są odpowiednikiem informacji — którą sprecyzowano w pierwszej części tej pracy — wyrażonej jedynie w terminach języka. Zbiory obiektów odpowiadające termom prymitywnym będziemy nazywali zbiorami elementarnymi (albo atomowymi). Wynika stąd, że językiem L_S można opisać tylko takie podzbiory zbioru obiektów, które są sumami zbiorów atomowych. A więc nie jest możliwe opisanie dowolnego podzbioru zbioru X w ogólnym przypadku.

Konsekwencje implementacyjne tego faktu są bardzo poważne. Wynika z niego, że wszystkie obiekty (ściślej ich opisy) można pogrupować w pamięci maszyny w zbiory atomowe (które łatwo obliczyć), a do uzyskania odpowiedzi na pytanie w systemie wystarczy dostęp jedynie do całych zbiorów atomowych, bez konieczności dostępu do dowolnych elementów zbioru X. Upraszcza to i przyspiesza znakomicie proces wyszukiwania.

W najogólniejszym zarysie system wyszukiwania może pracować następująco: pytanie postawione w systemie jest najpierw sprowadzone do postaci normalnej, a następnie, na podstawie odpowiednich algorytmów, nie podanych w tym artykule, można łatwo obliczyć adresy wszystkich zbiorów atomowych wchodzących w skład odpowiedzi na postawione pytanie, a następnie zbudować odpowiedź przez bezpośrednie wybranie obliczonych zbiorów atomowych.

Przy przeszukiwaniu dużych zbiorów X, system taki może okazać się bardzo szybkim. Daje on bowiem możliwość znalezienia bezpośrednio odpowiedzi, o którą chodzi użytkownikowi, bez konieczności szukania odpowiednich elementów odpowiedzi w całej pamięci.

Oczywiście w rzeczywistych systemach problem ten jest bardziej złożony. Tu przedstawiono jedynie schemat rozwiązania praktycznego. W przypadku niezbyt dużych zbiorów obiektów, proponowana metoda może być nieopłacalna. Sprowadzanie bowiem pytania do postaci normalnej, liczenie zbiorów atomowych może zająć więcej czasu aniżeli bezpośrednie odszukanie odpowiedzi w pamięci, dzięki metodzie list inwersyjnych czy nawet pełnego przeglądu.

5. Uwagi ogólne

Na marginesie przedstawionych tu rozważań powstaje wiele pytań natury bardziej ogólnej.

Pierwsze z nich to sprawa efektywności systemu. Chodzi o ściślejsze sformułowanie tego problemu i dokładniejsze jego zbadanie niż uczyniono to w tej pracy. Wydaje się celowe aby efektywność systemu wyszukiwania informacji mierzyć stosunkiem liczby obiektów faktycznie wchodzących w skład odpowiedzi na pytanie t do liczby obiektów koniecznych do odczytania z pamięci dla uzyskania odpowiedzi na pytanie t, tj.

$$E_t=\frac{|X_t|}{|X_{t'}|}$$

gdzie $|X_t|$ oznacza zbiór będący odpowiedzią na pytanie t, zaś $|X_{t'}|$ zbiór elementów faktycznie odczytywanych z pamięci dla uzyskania odpowiedzi t, natomiast $|X|$ oznacza liczbę elementów zbioru X. Wynika stąd, że metoda przeglądu zupełnego będzie miała dla każdego pytania najniższą efektywność, natomiast proponowana metoda będzie miała efektywność największą (równą 1). Przedstawiona propozycja ma jednakże różne wady. Najważniejsze jest to, że efektywność systemu odnosi się do konkretnego pytania. Można tę miarę efektywności rozszerzyć na dowolny zbiór pytań aby był on dostatecznie charakterystyczny dla określenia efektywności całego systemu. Sprawa ta nie jest prosta i konieczne są tu odpowiednie badania. Oczywiście można zamiast mówić o mierze efektywności systemu, wprowadzić hierarchię metod i wtedy stwierdzi się, że ta metoda jest bardziej efektywna, która dla każdego poszczególnego pytania posiada większą efektywność. Jednakże takie ujęcie może dawać wyniki niezgodne z wymaganiami praktyki, gdyż nie jesteśmy na ogół zainteresowani wszelkimi możliwymi pytaniami, a tylko tymi, które w jakimś sensie są najbardziej interesujące dla użytkownika. Scharakteryzowanie tej klasy pytań wydaje się jednakże sprawą trudną.

Następny problem stanowi sprawa dokładności systemu. Pojęcie to jest dość trudne do sformułowania, wydaje się jednak, iż można tu przyjąć następującą konwencję: możemy być zainteresowani wyszukaniem dowolnego podzbioru obiektu X. Jak wiemy w języku możemy zdefiniować tylko te podzbiory zbioru X, które są sumą zbiorów atomowych. Możemy więc wprowadzić pojęcie aproksymacji danego zbioru $X' \subset X$ poprzez sumę zbiorów atomowych; w ten sposób, że Y jest aproksymacją X', jeżeli Y jest sumą zbiorów atomowych zawierających X'. Możemy więc mówić również o minimalnym (tj. zawierającym, każdą inną aproksymację) zbiorze Y aproksymującym X', a wśród aproksymacji minimalnych możemy wybrać zbiór po-

siadający najmniejszą liczbę elementów. Aproksymację taką nazwiemy najlepszą. Różnica $|X'-\tilde{Y}|$ może być wtedy miarą dokładności odpowiedzi, gdzie $\tilde{Y}$ oznacza najlepszą aproksymację.

Jednak i tutaj sprawa nie jest całkiem prosta. Podobnie jak w przypadku efektywności można mówić w ten sposób jedynie o dokładności odpowiedzi w stosunku do ustalonego zbioru. Powstaje problem jak ten zbiór określić? W skonstruowanym przez nas języku informacyjnym możemy zdefiniować tylko te zbiory, które są sumami zbiorów atomowych. Do określenia zbioru, który ma być podstawą miary dokładności systemu musimy zastosować nowy, specjalny język, w przeciwnym wypadku taka definicja nie ma sensu. Jeżeli jednakże sprawa ma być sformułowana ściśle to ten nowy język musi być również określony formalnie, a więc będzie to język podobny do języka informacyjnego ze wszystkimi jego ograniczeniami, czyli również w ogólnym wypadku nie pozwalający na zdefiniowanie dowolnego podzbioru zbioru X. Możemy więc również, podobnie jak w przypadku efektywności, mówić o hierarchii systemów z punktu widzenia ich dokładności, wykazując, że dany system jest bardziej dokładny od pozostałych, dla każdego podzbioru obiektów.

Istnieje wiele innych problemów związanych z poruszoną tu problematyką, ważnych zarówno dla badań teoretycznych jak i praktycznej realizacji. Najważniejszym z nich wydaje się stworzenie teorii informacji, której punktem wyjścia byłyby fakty związane z wyszukiwaniem i magazynowaniem informacji — i która pozwalałaby na sprecyzowanie podstawowych pojęć oraz metod stosowanych w tej dziedzinie wiedzy.

002.513.5:681.3.004.14
025.4:681.3.04]STAIRS

System automatycznego wyszukiwania informacji — STAIRS w IDKKAP[1])

JACEK KAŹMIERCZAK
BARBARA MODZELEWSKA
ELŻBIETA WĘGŁOWSKA
Instytut Doskonalenia Kadr Kierowniczych Administracji Państwowej, Warszawa

Wstęp

W 1973 r. 13 instytucji: Biblioteka Narodowa, Biblioteka Główna Politechniki Krakowskiej, Biblioteka Główna SGPiS, Biblioteka Uniwersytetu Warszawskiego, Centrum Informatyki Handlu Zagranicznego, Główna Biblioteka Lekarska, Instytut Doskonalenia Kadr Kierowniczych Administracji Państwowej, Kancelaria Sejmu PRL, Ministerstwo Kultury i Sztuki — Zjednoczenie Księgarstwa, Ministerstwo Łączności, Polska Agencja Prasowa, Polski Komitet Normalizacji i Miar, Urząd Patentowy PRL, podjęło wspólny program badań (pod nazwą INFONET) nad przydatnością oprogramowania oraz sprzętu firm krajowych i zagranicznych dla potrzeb automatycznego wyszukiwania informacji.

Instytut Doskonalenia Kadr Kierowniczych Administracji Państwowej w Warszawie (Branżowy Ośrodek Informacji przy współpracy Zakładu Informatyki) prowadzi od 1974 r. badania testowe systemu STAIRS (STORAGE AND INFORMATION RETRIEVAL SYSTEM) uruchomionego na komputerze IBM SYSTEM 360/50, który został zainstalowany 10.09.1974 r. w IDKKAP.

Informacje ogólne

System STAIRS jest licencyjnym pakietem oprogramowania firmy IBM działającym na komputerach Systemu 360 i Systemu 370 w wersji OS (Operating System). Zadaniem tego systemu jest przechowywanie i zdalne konwersacyjne wyszukiwanie informacji o dowolnej tematyce, odpowiednio opracowanej i zapisanej na maszynowych nośnikach informacji.

1) Instytut Doskonalenia Kadr Kierowniczych Administracji Państwowej.